Monte Carlo Procedures for Variance Components

S. R. Searle, D. S. Robson, and Y. Y. Wang

BU-214-M April, 1966

ABSTRACT

Details are given for estimating, by Monte Carlo techniques, the probability density function of an estimated variance component. A distinction is made between simulation procedures and Monte Carlo procedures.

Biometrics Unit, Plant Breeding Department, Cornell University.

Monte Carlo Procedures for Variance Components

S. R. Searle, D. S. Robson, and Y. Y. Wang

BU-214-M April, 1966

Simulation and Monte Carlo

Computers are frequently used to simulate the drawing of sample data from hypothetical populations. Analyses can be made of these data to estimate statistics---variance components, for example---and from repeated (computer) sampling frequency distributions of the estimated variance components can be made, usually in the form of histograms. They represent estimates of the probability density functions of the variance component estimators. This method of estimating a density function can be referred to as estimation by simulation, to distinguish it from what is about to be described which shall be called estimation by Monte Carlo.

The procedure described above does nothing more than estimate a variance component from each of many samples of data and summarize the estimates so obtained. No use is made whatever of any information that is available concerning distributional properties of the estimator being used. Although the distribution function of most variance components estimators is unknown, most of these estimators do involve the "within" or "error" variance component estimator which, under conditions of normality, is usually distributed as χ^2. This information can be utilized to provide more efficient estimates of the density functions of the other variance components estimators than can be obtained from mere simulation. It is this method of estimating density functions, utilizing the χ^2-distribution

---

Biometrics Unit, Plant Breeding Department, Cornell University.

of the "within" mean square, that we call estimation by Monte Carlo. The general procedure is as follows.

Suppose an estimated variance component can be expressed as

$$\hat{\sigma}^2 = z + \lambda x$$

where z is a random variable (usually a linear function of sums of squares) having unknown distribution, where λ is a known constant and x is a χ^2 variable (a "within" mean square divided by its expectation). Then, apart from λ, the conditional variable $(\hat{\sigma}^2 | z)$ has a χ^2 distribution, and for any pre-assigned interval I_k, say, the probability $Pr\{\hat{\sigma}^2 | z \text{ lies in } I_k\}$ can be calculated. Through calculating these probabilities for a continuum of intervals, for each of many simulated z's (sums of squares), an estimate of the density function of $\hat{\sigma}^2$ is obtained. Details follow for the case of the between-groups variance component in a 1-way classification, with unequal subclass numbers.

1-way Classification

The analysis of variance for a 1-way classification, unbalanced data (unequal subclass numbers), can be expressed in familiar notation as follows:

Analysis of Variance for N Observations in c Classes.

Source	d.f.	Sum of Squares	Mean Squares	E (Mean Squares)
Classes	c-1	$\Sigma \bar{x}_{i.}^2/n_i - N\bar{x}_{..}^2 = SSB$	$MSB = SSB/(c-1)$	$k_o\sigma_a^2 + \sigma_e^2$
Residual	N-c	$\Sigma\Sigma x_{ij}^2 - \Sigma \bar{x}_{i.}^2/n_i = SSW$	$MSW = SSW/(N-c)$	σ_e^2
Total	N-1	$\Sigma\Sigma x_{ij}^2 - N\bar{x}_{..}^2$		

The expected values of the mean squares given in the last column are on the basis of a completely random model $x_{ij} = \mu + a_i + e_{ij}$, where μ is a general mean and a_i and e_{ij} are mutually independent random variables having variances σ_a^2 and

σ_e^2; the a_i and e_{ij} are also assumed to be normally distributed. The constant k_o is $k_o = (N - \Sigma n_i^2/N)/(c - 1)$.

Estimators of σ_a^2 and σ_e^2 can be had from Henderson's (1953) Method 1, of equating the observed mean squares to their expected values. This gives

$$\hat{\sigma}_e^2 = MSW = \frac{SSW}{N-c} = \frac{\sigma_e^2}{N-c} \cdot \frac{SSW}{\sigma_e^2} = \frac{\sigma_e^2}{N-c}\, x \qquad \text{- - - (1)}$$

where $x = SSW/\sigma_e^2$ is a random variable distributed as χ^2 with N-c degrees of freedom; and

$$\hat{\sigma}_a^2 = \frac{MSB - MSW}{k_o} = \frac{SSB}{k_o(c-1)} - \frac{\sigma_e^2}{k_o(N-c)} \frac{SSW}{\sigma_e^2} = \lambda_1 SSB - \lambda_2 x \qquad \text{- - - (2)}$$

where

$$\lambda_1 = \frac{1}{k_o(c-1)} = \frac{N}{N^2 - \Sigma n_i^2} \quad \text{and} \quad \lambda_2 = \frac{\sigma_e^2}{k_o(N-c)} = \frac{(c-1)\lambda_1 \sigma_e^2}{N-c} \,. \qquad \text{- - - (3)}$$

We note that the random variable x, distributed as χ^2_{N-c}, is distributed independently of SSB.

From (2) it is evident that the conditional variable $(\hat{\sigma}_a^2 | SSB)$ is distributed as $-\lambda_2 x$. Now suppose I_k is an interval (a_{k-1}, a_k) on the real line, with $a_{k-1} < a_k$. Then the probability that $\hat{\sigma}_a^2 | SSB$ lies within this interval is

$$P_k = Pr\{\hat{\sigma}_a^2 | SSB \in I_k\} = Pr\{\lambda_1 SSB - \lambda_2 x \in I_k\}$$

$$= Pr\{a_{k-1} < \lambda_1 SSB - \lambda_2 x \leq a_k\}$$

$$= Pr\left\{\frac{\lambda_1 SSB - a_k}{\lambda_2} \leq x < \frac{\lambda_1 SSB - a_{k-1}}{\lambda_2}\right\} \qquad \text{- - - (4)}$$

$$= Pr\{L_k \leq x < L_{k-1}\}, \qquad \text{- - - (5)}$$

where

$$L_k = \frac{\lambda_1 SSB - a_k}{\lambda_2} \quad \text{and} \quad L_{k-1} = \frac{\lambda_1 SSB - a_{k-1}}{\lambda_2} \,. \qquad \text{- - - (6)}$$

With x being distributed as χ^2, let $Q(z)$ be the probability that x exceeds z. Then (5) is

$$P_k = Q(L_k) - Q(L_{k-1})$$

where L_k and L_{k-1} are as in (5).

Now consider a situation in which, for given N and c, a given set of n_i values, and for given values of σ_a^2 and σ_e^2, we simulate a series of SSB values. λ_1 and λ_2 in (3) are then known, and on defining I_k, a_k and a_{k-1} are also known. Thus from (6), L_k and L_{k-1} can be calculated and so, with x being a χ^2_{N-c} variable, tables of the χ^2-distribution (or some other procedure) can be used to calculate $\Pr\{L_k \leq x < L_{k-1}\} = P_k = \Pr\{\hat{\sigma}^2 | \text{SSB} \in I_k\}$ as in (5). And this can be done for every I_k, for each SSB that is simulated; i.e. for each simulated SSB $\Pr\{\hat{\sigma}_a^2 | \text{SSB} \in I_k\}$ can be calculated for every I_k. And after simulating a large number of SSB's, M of them say, with M = 1000 or 2000 perhaps, the average value of the M probabilities calculated for each interval I_k can be found:

$$\bar{P}_k = \frac{1}{M} \sum_{i=1}^{M} [Q(L_{k,i}) - Q(L_{k-1,i})]$$

where $L_{k-1,i}$ and $L_{k,i}$ are the values of L_{k-1} and L_k given by (6) for the i'th simulated value of SSB. This average, $\bar{P}_k$, is an estimate of

$$E_{SSB}\left[\Pr\{\hat{\sigma}^2 | \text{SSB} \in I_k\}\right] = \Pr\{\hat{\sigma}_a^2 \in I_k\},$$

where E_{SSB} denotes expectation with respect to sampling of the SSB. Viewed over the whole continuum of intervals the estimates $\bar{P}_k$ constitute a histogram estimator of the density function of $\hat{\sigma}_a^2$.

Evaluation of Probabilities

Consider the variable x in (4). It has a χ^2 distribution and so cannot take negative values. But for some simulated SSB suppose that k' is such that

$$a_{k'-1} < \lambda_1 SSB < a_{k'} \cdot \qquad \text{- - - (7)}$$

Then, from (6), $L_{k'} = (\lambda_1 SSB - a_{k'})/\lambda_2$ is negative. And, because $\cdots a_{k'+2} > a_{k'+1} > a_{k'}$, $a_k > a_{k'}$ for $k > k'$, and so for these values of k all values of L_k are negative. Since x cannot be negative we define P_k as zero in these cases; i.e. for $k > k'$, $P_k = 0$. It can also be seen from (7) that for k' defined therein $L_{k'-1} = (\lambda_1 SSB - a_{k'-1})/\lambda_2$ is positive. Hence

$$P_{k'} = Pr\{L_{k'} \leq x < L_{k'-1}\}$$

is redefined as

$$P_{k'} = Pr\{0 \leq x < L_{k'-1}\} \cdot$$

Also from (7), because $a_{k'} > a_{k'-1} > a_{k'-2} \cdots$, all values of L_k for $k < k'$ are positive. Hence for $k < k'$ the value of P_k is calculated as it stands.

The calculation of P_k over the whole range of intervals I_k therefore falls into three phases, depending on the value of k relative to k':

$$\begin{aligned} P_k &= 0 && \text{for } k > k' \\ &= Pr\{0 \leq x < L_{k'-1}\} && \text{for } k = k' \\ &= Pr\{L_k \leq x < L_{k-1}\} && \text{for } 0 \leq k < k' \end{aligned} \qquad \text{- - - (8)}$$

It can be noted in passing that the definition of k' given by (7) is that interval $I_{k'}$ which contains $\lambda_1 SSB$.

Choice of Intervals

As shown in (3), λ_2 is linear in σ_e^2. So is SSB, for it is simulated through simulating cell means $\bar{x}_i$, i = 1, 2, ···, c as

$$\begin{aligned} \bar{x}_i &= \alpha_i \sigma_a + \gamma_i \sigma_e / \sqrt{n_i} \\ &= \sigma_e \left[\alpha_i (\sigma_a / \sigma_e) + \gamma_i / \sqrt{n_i} \right] \end{aligned} \qquad \text{- - - (9)}$$

where α_i and γ_i are independently simulated standardized normal variates. σ_e^2 is thus a factor of SSB. If, therefore, a_k and a_{k-1} are chosen in such a way that they too contain σ_e^2 as a factor, σ_e^2 will cancel out from L_k and L_{k-1} as given in (6). This can be achieved by at least one method of choosing the intervals I_k that utilizes the sampling standard error of $\hat{\sigma}_a^2$, a method that is also intrinsically appealing for other reasons. It proceeds thus.

Suppose $s\sigma_e^2$ denotes the sampling standard error of $\hat{\sigma}_a^2$. Then by adaptation (and minor correction) of the formula given in Searle (1956)

$$s = \left[\sqrt{\text{var}(\hat{\sigma}_a^2)} \right] / \sigma_e^2$$

$$= \sqrt{2}\left[\frac{(N-1)(c-1)N^2}{(N-c)(N^2-S_2)^2} + \frac{\sigma_a^2}{\sigma_e^2} \cdot \frac{2N}{N^2-S_2} + \left(\frac{\sigma_a^2}{\sigma_e^2}\right)^2 \frac{N^2S_2 + S_2^2 - 2NS_3}{(N^2 - S_2)^2} \right]^{\frac{1}{2}} \qquad (10)$$

where $S_2 = \Sigma n_i^2$ and $S_3 = \Sigma n_i^3$. By definition, s is the standard error of $(\hat{\sigma}_a^2/\sigma_e^2)$, and if a_k is now taken in the general form

$$\left(\frac{\sigma_a^2}{\sigma_e^2} + p_k s\right)\sigma_e^2 \qquad \text{- - - (11)}$$

for p_k being some suitably assigned constant, then σ_e^2 will cancel from L_k and L_{k-1}; operationally this is equivalent to putting $\sigma_e^2 = 1$ in (3), (9), (10) and (11), thereupon interpreting σ_a^2 as the ratio σ_a^2/σ_e^2. With this interpretation the intervals will be

$$\begin{array}{llll} I_0 & -\infty & \text{to } \sigma_a^2 - p_0 s = a_0 ; \\ I_1 & a_0 & \text{to } \sigma^2 - p_1 s = a_1 ; \\ \vdots & & \vdots & \qquad \text{- - - (12)} \\ I_m & a_{m-1} & \text{to } \sigma_a^2 - p_m s = a_m , \\ I_{m+1} & a_m & \text{to } +\infty . \end{array}$$

where the constants $p_0, p_1, \ldots, p_m$ are chosen in some suitable manner. Presumably the intervals might be symmetric about σ_a^2 - but from evidence gained in simulating $\hat{\sigma}_a^2$ and plotting its frequency distribution about σ_a^2 it might be more appropriate to take $p_0 = 2$ and $p_m = 2$, 3, 4 or 5, using approximately 100 intervals if $p_m = 5$, and 60 if $p_m = 2$.

Advantage of procedure

In the simulation procedure of just plotting a frequency distribution of $\hat{\sigma}_a^2$, each simulated $\hat{\sigma}^2$ contributes information to only one interval of the histogram that ultimately estimates the frequency distribution, namely the interval in which it falls. But in the Monte Carlo method just outlined each simulated SSB will contribute information to every interval. This will yield a more reliable estimate of the density function than would the same amount of simulation used in merely plotting a frequency function of $\hat{\sigma}_a^2$. In effect, this Monte Carlo procedure removes the variation due to the χ^2 variable that forms part of the estimate $\hat{\sigma}_a^2$. And in particular, the tails of the distribution should be better estimated.

Computing

For each set of n_i's (n-pattern) decided upon:

k_0 ;

λ_1 and λ_2 from (3), using $\sigma_e^2 = 1$;

s as in (10), with $\sigma_e^2 = 1$.

For each σ_a^2, and for the constants $p_0, p_1, \ldots, p_m$:

the m + 2 intervals in (12).

For each of the M simulations:

$\bar{x}_1, \ldots, \bar{x}_c$ as in (9), using $\sigma_e = 1$

SSB as in the analysis of variance

a'_k as in (7)

the k' + 1 probabilities given in (8).

After M simulations:

divide by M the total probability accumulated in each of the m + 2 intervals.

The probabilities

$$\Pr\{L_k \leq x < L_{k-1}\}$$

required in (8) can be computed as

$$Q(L_k) - Q(L_{k-1}) = \sum_{j=0}^{\frac{1}{2}(N-c)} \frac{e^{-\frac{1}{2}L_k}(\frac{1}{2}L_k)^j - e^{-\frac{1}{2}L_{k-1}}(\frac{1}{2}L_{k-1})^j}{j!}$$

(Formulae 26.4.21, Nat. Bur. of Standards, 1964).

This holds good only for N - c an even number.

References

Henderson, C. R. (1953). Estimation of variance and covariance components. Biometrics 9, 226-252.

National Bureau of Standards, (1964). Handbook of Mathematical Functions, Abramowitz and Stegun, Editors.

Searle, S. R. (1956). Matrix methods in components of variance and covariance. Ann. Math. Stat., 27, 737-748.